View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Elsevier - Publisher Connector

Redifusión de catálogos bibliográficos en MARC-XML

Manuel Blázquez Ochando *

Artículo recibido:
21 de mayo de 2013.
Artículo aceptado:
7 de agosto de 2013.

Resumen

La técnica de redifusión de noticias habitualmente empleada en los medios de comunicación social puede ser utilizada en el contexto de los catálogos bibliográficos, modificando su formato de codificación RSS por MARC-XML. Las ventajas que se desprenden de este uso son una mayor difusión de los fondos, la posibilidad de compartir los catálogos con terceras bibliotecas y permitirles a los usuarios la descarga y lectura de éstos. Para lograrlo se han llevado a cabo diversas pruebas que miden el tiempo de creación y recuperación de tales colecciones bibliográficas. Por otro lado se determina qué tipo de programas se necesitan para operar con dichos archivos y cuál es su funcionamiento. Como

* Universidad Complutense de Madrid, España. manuel.blazquez@pdi.ucm.es

resultado de estas experiencias se concluye que es factible generar, transmitir y recuperar catálogos bibliográficos mediante técnicas inspiradas en la sindicación de contenidos.

Palabras clave: Redifusión de contenidos; MARC-XML; Transmisión de registros bibliográficos; Canales de sindicación bibliográficos.

Abstract

Rebroadcasting of bibliographic catalogues in MARC-XML format

Manuel Blázquez-Ochando

By using MARC-XML to modify the RSS code format, the technique habitually used by the media in rebroadcasting news can also be used for bibliographical catalogues. Among other things, this procedure offers the advantages of improved dissemination of contents, sharing catalogues with other libraries, and allowing user downloads of catalogues. Researchers performed an array of trials to measure the building and recovery times for such bibliographical collections, while determining the sort of applications and functions needed to control these files. These experiences allow researchers to conclude that it is possible to generate, transmit and retrieve bibliographical catalogues using content syndication practices and methods.

Keywords: Content syndication; MARC-XML; Transmission of bibliographic records; Bibliographic syndication channels.

Introducción

Los catálogos bibliográficos en línea constituyen una herramienta de primera necesidad en cualquier unidad de información y documentación. Los servicios comúnmente ofrecidos al usuario abarcan la exportación de los registros consultados destinados a programas de gestión bibliográfica, su etiquetado social, su referenciación en nuevas obras científicas, su posterior

consulta, acceso y recuperación a texto completo. Uno de los retos de la constante evolución de los catálogos es su mayor difusión, hasta el punto de ser enteramente compartidos con los usuarios de una manera distribuida y libre. Este objetivo puede ser alcanzado mediante la conversión de los catálogos bibliográficos en formato MARC-XML y su tratamiento adecuado mediante programas *parser* o analizadores de estructuras basadas en XML.

La transferencia de registros bibliográficos vía HTTP por medio de archivos MARC-XML puede basarse en las técnicas de sindicación o asociación de contenidos, tal como sugiere Blázquez Ochando (2010: 228-392). En esta tesis doctoral se postula que las técnicas de sindicación de contenidos, inicialmente utilizadas para la redifusión de noticias en los medios de comunicación social, podrían utilizarse para la transmisión y recuperación de catálogos bibliográficos de manera completa o parcial, tal como ya se puede comprobar en la Biblioteca Digital de Munich, un año después de su publicación (Münchener Digitalisierungszentrum Digitale Bibliothek, 2011). Otra experiencia avanzada que demuestra el interés de los centros de información por incorporar el estándar MARC-XML es la iniciativa del catálogo de tesis doctorales de la Universidad de Sevilla, que posibilita la exportación y descarga libre de sus registros en dicho formato (Universidad de Sevilla, 2011).

En el campo de la Documentación, las aplicaciones más habituales de la sindicación consisten en la creación de canales de información generales, el establecimiento de servicios de alertas bibliográficas (*ANU Library: new titles*, 2011), la redifusión de artículos y contenidos de revistas científicas (Rodríguez Gairín *et al*., 2006) o la agrupación de consultas en canales de sindicación personalizados (PUBMED, 2011; Dolan, 2011) en las que la rama biosanitaria de la Documentación es mucho más activa.

En este artículo se analiza cómo generar catálogos bibliográficos en formato MARC-XML para posteriormente recuperarlos por medio de programas *parser* similares a los empleados por los lectores de canales de sindicación. Para verificar este proceso se aportará una prueba real sobre la viabilidad de la transmisión de catálogos bibliográficos en red y una prueba de ejecución de tales programas en la plataforma habilitada al efecto *OrangeUP* (Blázquez Ochando, 2011).

Generación de catálogos en marc-xml

Generar catálogos en formato MARC-XML (Library of Congress, 2011) requiere de la disponibilidad de los registros bibliográficos en base de datos

para su completa gestión y tratamiento. De no ser así será necesario migrar la información. Un método para conseguir transferir el catálogo bibliográfico es la exportación del mismo a formato CSV, opción común en la mayoría de los gestores bibliográficos y bibliotecarios. Para este estudio se han compuesto diversas colecciones que abarcan desde los mil registros hasta el millón, a partir del catálogo bibliográfico de la Library of Congress (Blázquez Ochando, 2010: 299), como se muestra en la *Tabla 1.*

Tabla 1. Características de las colecciones bibliográficas probadas

Colección	Tamaño en disco	Núm. de registros
1000_reg	0.77	1 001
5000_reg	2.68	5 002
10000_reg	5.05	10 004
25000_reg	13.33	25 008
50000_reg	28.34	50 036
100000_reg	54.95	100 054
250000_reg	144.00	250 146
500000_reg	280.49	500 309
1000000_reg	561.39	1 000 039

Este paso previo hace posible que, mediante un programa de exportación desarrollado en PHP (Blázquez Ochando, 2010: 268-271), se genere un catálogo en MARC-XML correspondiente al catálogo bibliográfico inicial. Para ello se reproduce la estructura básica del registro, el nodo *record* y sus dependientes, tantas veces como ejemplares y volúmenes tenga el fondo en cuestión (véase *Tabla 2).*

Tabla 2. Registro modelo utilizado

```
<record>

  <controlfield tag='001'>Nº Control interno</controlfield>
  <controlfield tag='003'>Nº identificación del documento</controlfield>

  <datafield tag='017' ind1='' ind2=''>
    <subfield code='a'>Depósito legal o Copyright</subfield>
  </datafield>

  <datafield tag='020' ind1='' ind2=''>
    <subfield code='a'>ISBN</subfield>
  </datafield>

  <datafield tag='022' ind1='0' ind2=''>
    <subfield code='a'>ISSN</subfield>
  </datafield>
```

```
<datafield tag='035' ind1='' ind2=''>
  <subfield code='a'>Número de Control del Sistema</subfield>
</datafield>

<datafield tag='041' ind1='0' ind2=''>
  <subfield code='a'>Código del idioma del documento original</subfield>
</datafield>

<datafield tag='043' ind1='' ind2=''>
  <subfield code='c'>Código geográfico del documento original</subfield>
</datafield>

<datafield tag='082' ind1='' ind2=''>
  <subfield code='a'>Clasificación Decimal Dewey</subfield>
</datafield>

<datafield tag='100' ind1='1' ind2=''>
  <subfield code='a'>Autor personal</subfield>
</datafield>

<datafield tag='245' ind1='1' ind2=''>
  <subfield code='a'>Área de título</subfield>
  <subfield code='b'>Subtítulo</subfield>
  <subfield code='c'>Mención de responsabilidad</subfield>
</datafield>

<datafield tag='250' ind1='' ind2=''>
  <subfield code='a'>Nº de edición</subfield>
  <subfield code='b'>Mención de edición</subfield>
</datafield>

<datafield tag='260' ind1='' ind2=''>
  <subfield code='a'>Lugar de publicación</subfield>
  <subfield code='b'>Editorial</subfield>
  <subfield code='c'>Año de publicación</subfield>
</datafield>

<datafield tag='300' ind1='' ind2=''>
  <subfield code='a'>Área de descripción física</subfield>
</datafield>

<datafield tag='310' ind1='' ind2=''>
  <subfield code='a'>Periodicidad</subfield>
</datafield>

<datafield tag='490' ind1='0' ind2=''>
  <subfield code='a'>Serie o colección</subfield>
  <subfield code='v'>Nº de serie o colección</subfield>
</datafield>

<datafield tag='500' ind1='' ind2=''>
```

```
    <subfield code='a'>Área de notas</subfield>
  </datafield>

  <datafield tag='654' ind1='0' ind2=''>
    <subfield code='a'>Temática del documento</subfield>
  </datafield>

</record>
```

El factor que interviene en el proceso anteriormente descrito es el volumen de la codificación de los registros bibliográficos y su extensión descriptiva. Sobre la extensión del catálogo ha de advertirse que el tamaño de las colecciones a partir de 5 000 registros supera los 2 MB por fichero. Este detalle, que también ha sido contrastado y verificado *a posteriori* por la empresa especializada IndexData (Schafroth, 2010), implica que generar el catálogo correspondiente en un solo archivo XML multiplica el tamaño, ya que incluye caracteres destinados a su etiquetado y esto hace difícil su tratamiento, visualización y recuperación posteriores, como se venía señalando anteriormente (Blázquez Ochando, 2010: 257-258).

La solución a tal problema es crear un archivo XML cada 1 000 registros debido a que su tamaño rara vez supera 1 MB, lo que lo hace más sencillo de gestionar. Derivadas de esta solución, las grandes colecciones tienden a generar gran cantidad de archivos XML, lo que dificulta el acceso a la información del catálogo. Esta contraindicación se puede solucionar mediante el empleo de un archivo OPML que los agrupe, tal como indica la motivación de sus especificaciones (Winer, 2007). De esta forma es posible recuperar los catálogos completos en bloque (Blázquez Ochando, 2010: 278).

Recuperación de catálogos en marc-xml

El método de recuperación de catálogos en formato MARC-XML hace uso de programas de tipo *parser* capaces de analizar la estructura del archivo XML y de volcar la información para su aprovechamiento, ya sea su representación, filtrado o recuperación para el almacenamiento en bases de datos. Se trata en definitiva del proceso que cualquier agregador o lector de sindicación lleva a cabo de manera habitual, trasladado al contexto de los formatos bibliográficos, de mayor interés para la Documentación.

El ejemplo que se expone *a continuación* en la *Tabla 3* es un programa *parser* elaborado en PHP capaz de leer y recuperar un catálogo bibliográfico codificado en MARC-XML, como el expuesto en la *Tabla 4*. La clave de funcionamiento se

halla en la función *simplexml_load_file()* disponible en PHP GROUP (2011a). Tal y como se especifica, interpreta cualquier archivo basado en XML y lo convierte en un objeto que puede ser accesible en todos sus elementos por medio de DOM (PHP GROUP, 2011b).

Tabla 3. Modelo de selección de campos con XPath

```
<?php

$feed = "catalogo.xml";
$xml = simplexml_load_file($feed);

for($i=0; $xml->record[$i]; $i++) {

    // Campos de control
    $tag001 = $xml->record[$i]->controlfield[0];
    $tag005 = $xml->record[$i]->controlfield[1];

    // Entradas principales
    $tag100a = $xml->record[$i]->datafield[7]->subfield[0];

   // Área de título y mención de responsabilidad
    $tag245a = $xml->record[$i]->datafield[8]->subfield[0];
    $tag245b = $xml->record[$i]->datafield[8]->subfield[1];
    $tag245c = $xml->record[$i]->datafield[8]->subfield[2];

    // Área de publicación
    $tag260a = $xml->record[$i]->datafield[10]->subfield[0];
    $tag260b = $xml->record[$i]->datafield[10]->subfield[1];
    $tag260c = $xml->record[$i]->datafield[10]->subfield[2];

}

?>
```

Para verificar este extremo, una vez cargado el catálogo en la variable *$xml*, se puede imprimir en pantalla empleando la función *print_r($xml)* y obtener un resultado similar al expuesto en la *Figura 1*.

Tabla 4. Fragmento del array de datos recuperados del catálogo bibliográfico en MARC-XML

```
SimpleXMLElement Object (
[record] => SimpleXMLElement Object (
   [leader] => cabecera[controlfield] => Array (
      [0] => número de control
      [1] => identificador del número de control
```

```
        [2] => fecha y hora de la última actualización )
    [datafield] => Array (
        [0] => SimpleXMLElement Object (
            [@attributes] => Array (
                [tag] => 010
                [ind1] =>
                [ind2] => )
        [subfield] => número de control de la biblioteca del congreso )
            [1] => SimpleXMLElement Object (
            [@attributes] => Array ...
```

Para recuperar los datos de cada registro bibliográfico hay que recorrer todos los nodos <*record*> del catálogo en MARC-XML. Esta tarea es llevada a cabo por un bucle *for* cuyo límite de ejecución es precisamente el número total de entradas del archivo XML a tratar. Dentro de su ejecución se puede distinguir cómo se seleccionan las etiquetas codificadas en formato MARC, almacenadas en variables que llevan su nombre exacto. Por ejemplo: la etiqueta *100$a*, que representa al autor principal, se almacena en la variable *$tag100a* y corresponde al nodo <*datafield*> dispuesto en la posición número 7, cuyo valor es almacenado a su vez en la etiqueta <*subfield*>. Obsérvese que para alcanzar el valor alojado en estas etiquetas es necesario indicar la ruta de selección de principio a fin, partiendo en todo caso de la variable matriz *$xml,* que como se ha explicado anteriormente, alberga el contenido de todo el catálogo.

Pruebas con catálogos bibliográficos marc-xml

Para verificar el funcionamiento del método de generación y recuperación de catálogos bibliográficos en formato MARC-XML, se desarrolló el programa SYNC (Blázquez Ochando, 2010: 235-310), que permite llevar a cabo tales operaciones y efectuar una medición de los tiempos de ejecución, así como determinar en todo caso el buen término o no del experimento. Los resultados obtenidos en la *Tabla 5* reflejan que el proceso de generación automática de los catálogos supera los 15 minutos en el caso de la colección del millón de registros.

Tabla 5. Tiempo de creación de catálogos MARC-XML

Colección	Tiempo (segundos)
1000_reg	0.24
5000_reg	0.81
10000_reg	1.75

25000_reg	4.82
50000_reg	11.78
100000_reg	26.68
250000_reg	105.28
500000_reg	321.08
1000000_reg	1095.83

Aun así, los valores obtenidos con colecciones relativamente amplias como la de 50 000 registros se crean en poco más de 10 segundos. Estos datos se contraponen a los obtenidos en el proceso de recuperación de los catálogos. Ello es lógico dado que la operación de volcado de la información sólo requiere de lectura y escritura desde una fuente de información ya conocida, la base de datos.

Tabla 6. Tiempos de importación de catálogos MARC-XML

Colección	Tiempo (segundos)
1000_reg	1.68
5000_reg	8.36
10000_reg	16.85
25000_reg	42.63
50000_reg	92.88
100000_reg	184.64
250000_reg	510.92
500000_reg	1034.99
1000000_reg	2857.61

Cuando el proceso es a la inversa, el programa *parser*, tiene que leer el archivo XML, generar un objeto accesible por medio de DOM, seleccionar la ruta en la que se encuentra la información, presentarla en pantalla y finalmente insertarla en la base de datos. Como se puede apreciar en la *Tabla 6*, esta rutina duplica ampliamente el tiempo de ejecución y dificulta enormemente el trabajo con grandes colecciones, especialmente aquellas cercanas a 100 000 registros, con tiempos superiores a los 3 minutos.

Prueba de edición y publicación de catálogos marc-xml y rss

Con el objetivo de determinar las diferencias en el desarrollo de catálogos bibliográficos en formato MARC-XML y RSS, se propone una prueba de edición

y publicación que puede ser consultada en línea.[1] Su funcionamiento responde a una cadena de procesos claramente delimitados, (*Figura 1).*

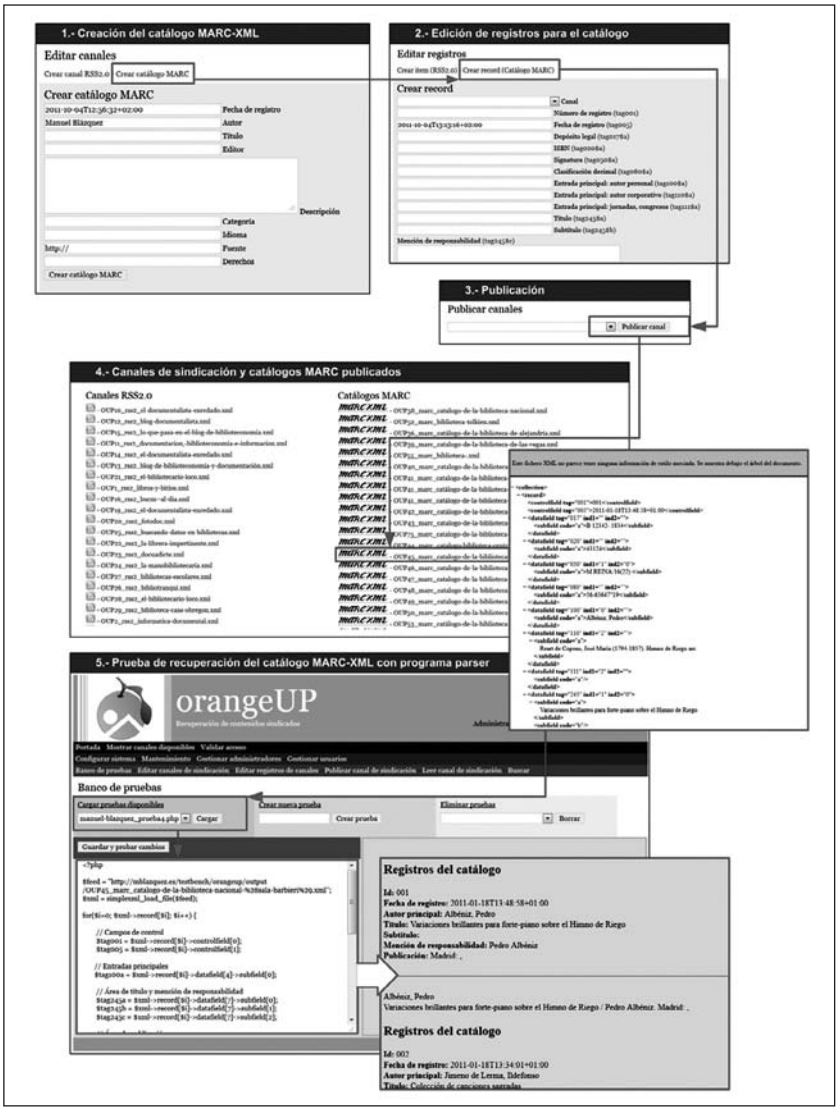

Figura 1. Cadena de procesos en la edición y publicación de catálogos bibliográficos: http://www.mblazquez.es/documents/orangeup-process.png

1 Véase programa de demostración OrangeUP, disponible en: http://www.mblazquez.es/testbench/orangeup/

El sistema *OrangeUP* ha sido desarrollado ex profeso para la gestión de canales de sindicación y para demostrar que independientemente del formato que se utilice para describir los registros bibliográficos o los contenidos informativos, todos los formatos basados en XML tienen las mismas propiedades de transmisión, compartición, edición, publicación y lectura. En la *Figura 1* se observa, en los primeros pasos, la creación de los catálogos bibliográficos en formato MARC-XML y RSS indistintamente, mediante el mismo método de edición y formularios. Los registros bibliográficos pueden ser editados conforme a las normas de descripción bibliográfica MARC21, manteniendo su codificación clave. A cada registro bibliográfico se le asigna un canal de sindicación bibliográfico. En todos los casos los registros creados son practicables y modificables, esto es, su meta-información, campos de descripción bibliográfica, categoría, clasificación, áreas de título y mención de responsabilidad entre otros. Durante este proceso, el programa almacena la información en base de datos MySQL para su posterior publicación y codificación en archivos cuyo formato MARC-XML o RSS será elegido por el usuario.

A continuación el usuario registrado puede utilizar la función *banco de pruebas*. Se trata de un editor de códigos para probar programas de tipo *parser*. Ello permite poner a prueba códigos como el reseñado en la *Tabla 3* y ejecutarlos de forma tal que se observen los resultados de los mismos sin salir de la pantalla. El resultado de este proceso es una visualización completa de todos los registros bibliográficos descritos en el catálogo al igual que lo haría un lector de canales de sindicación con sus respectivos ítems. Es por tanto innegable el paralelismo en la técnica de redifusión de contenidos o sindicación de contenidos y la técnica de redifusión de catálogos bibliográficos, salvando la diferencia de la codificación y de la popularidad de un formato RSS frente a MARC-XML. Para observar el proceso de edición y publicación del programa, véase el video original de la demostración con *OrangeUP*, disponible en http://youtu.be/kS2WiXuRFpM

Conclusiones

Es posible recuperar catálogos bibliográficos en formato MARC-XML utilizando programas *parser* análogos a los empleados para la lectura de canales de sindicación. Ello implica que los catálogos bibliográficos pueden ser compartidos entre bibliotecas empleando la metodología anteriormente citada.

Los tiempos de lectura y recuperación de los catálogos bibliográficos son superiores a los de creación debido a dos factores clave: por un lado la codificación del formato MARC-XML, bastante extensa en comparación con RSS, y por otro la extensión de las descripciones de los documentos del catálogo.

Cada vez se empiezan a aplicar más las técnicas de sindicación de catálogos bibliográficos, como en la Biblioteca Digital de Munich; en otros casos, ya se permite la exportación de los registros bibliográficos en formato MARC-XML para ser compartidos y reutilizados por terceros, como en el catálogo de tesis doctorales de la Universidad de Sevilla. Ello parece indicar el comienzo de la implementación de tales sistemas, así como su interés y experimentación en el entorno bibliotecario y documental.

Bibliografía

ANU Library: new titles (2011), disponible en: http://anulib.anu.edu.au/about/news/newbooks/ (Fecha de consulta: 12 de septiembre del 2011).

Blázquez Ochando, M. (2010), *Aplicaciones de la sindicación para la gestión de catálogos bibliográficos,* Madrid: Universidad Complutense.

—— (2011), *OrangeUp*, disponible en: http://mblazquez.es/testbench/orangeup/ (Fecha de consulta: 17 de marzo del 2011).

Dolan, F. (2011), *MedWorm,* disponible en: http://www.medworm.com/ (Fecha de consulta: 15 de marzo del 2011).

Library of Congress (2011), *MARC21 XML Schema*, disponible en: http://www.loc.gov/standards/marcxml/ (Fecha de consulta: 17 de septiembre del 2011).

Münchener Digitalisierungszentrum Digitale Bibliothek (2011), disponible en: http://www.digital-collections.de/index.html?c=rss&l=en (Fecha de consulta: 12 de septiembre del 2011).

PHP GROUP (2011a), *simplexml_load_file,* disponible en: http://php.net/manual/es/function.simplexml-load-file.php (Fecha de consulta: 26 de septiembre del 2011.

—— (2011b), *Document Object Model,* disponible en: http://php.net/manual/es/book.dom.php (Fecha de consulta: 26 de septiembre del 2011).

PUBMED (2011), disponible en: http://www.ncbi.nlm.nih.gov/pubmed (Fecha de consulta: 15 de marzo del 2011).

Rodríguez Gairín, J.M. *et al.* (2006), "Sindicación de contenidos en un portal de revistas: Temaria", en *El Profesional de la Información*, 15 (3), pp. 214-221.

Schafroth, D. (2010), *Turbomarc, faster XML for MARC records*, disponible en: https://www.indexdata.com/blog/2010/05/turbomarc-faster-xml-marc-records (Fecha de consulta: 18 de marzo del 2011).

Universidad de Sevilla (2011), *Tesis Doctorales: fondos digitalizados,* disponible en: http://fondosdigitales.us.es/tesis/ (Fecha de consulta: 17 de marzo del 2011).

Winer, D. (2007), *OPML 2.0*, disponible en: http://www.opml.org/spec2 (Fecha de consulta: 17 de marzo del 2011).